(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) EP 2 437 517 A1

(12) EUROPEAN PATENT APPLICATION

(43) Date of publication:
04.04.2012 Bulletin 2012/14

(51) Int Cl.:
H04R 3/00 *(2006.01)* ***G10L 21/02*** *(2006.01)*

(21) Application number: **10275102.1**

(22) Date of filing: **30.09.2010**

(84) Designated Contracting States:
AL AT BE BG CH CY CZ DE DK EE ES FI FR GB GR HR HU IE IS IT LI LT LU LV MC MK MT NL NO PL PT RO SE SI SK SM TR
Designated Extension States:
BA ME RS

(71) Applicant: **NXP B.V.**
5656 AG Eindhoven (NL)

(72) Inventors:
- **Van Waterschoot, Toon**
  Redhill
  Surrey RH1 1DL (GB)
- **Tirry, Wouter**
  Redhill
  Surrey RH1 1DL (GB)
- **Moonen, Marc**
  Redhill
  Surrey RH1 1DL (GB)

(74) Representative: **Williamson, Paul Lewis**
NXP Semiconductors UK Ltd.
Intellectual Property Department
Betchworth House
57-65 Station Road
Redhill
Surrey RH1 1DL (GB)

Remarks:
Amended claims in accordance with Rule 137(2) EPC.

(54) Sound scene manipulation

(57) An audio-processing device. The device comprises: an audio input, for receiving one or more audio signals detected at respective microphones. Each of the audio signals comprises a mixture of a plurality of components, each component corresponding to a sound source. The device also comprises a control input, for receiving, for each sound source, a desired gain factor associated with the source, by which it is desired to amplify the corresponding component. The device further comprises an auxiliary signal generator, adapted to generate at least one auxiliary signal from the one or more audio signals, the at least one auxiliary signal comprising a different mixture of the components as compared with a reference one of the one or more audio signals; and a scaling coefficient calculator, adapted to calculate a set of scaling coefficients in dependence upon the desired gain factors and upon parameters of the different mixture, each scaling coefficient associated with one of the at least one auxiliary signal and optionally the reference audio signal. It also comprises an audio synthesis unit, adapted to synthesize an output audio signal by applying the scaling coefficients to the at least one auxiliary signal and optionally the reference audio signal and to combine the results. The scaling coefficients are calculated from the desired gain factors and the parameters of the different mixture such that the synthesized output signal provides the desired gain factor for each component.

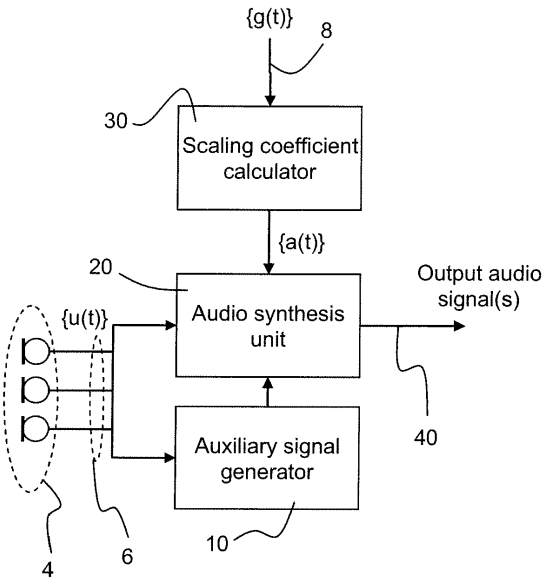

FIG 1

EP 2 437 517 A1

Description

[0001] This invention relates to manipulation of a sound scene comprising multiple sound sources. It is particularly relevant in the case of simultaneous recording of audio by multiple microphones.

[0002] Most existing sound scene manipulation methods operate in a two-stage fashion: in a first stage, the individual sound sources are extracted from one or more microphone recordings; and in a second stage, the separated sound sources are recombined according to the desired sound scene manipulation. When the manipulation consists of a change in the desired level of the individual sound sources (which is commonly the case), the second stage is trivial, once the first stage has been executed. Indeed, the recombination in the second stage then reduces to a simple linear combination of the separated sound sources obtained from the first stage. Unfortunately, the extraction of the individual sound sources from the recorded microphone signal(s) is a difficult problem, on which a lot of research effort has been spent. Broadly speaking, the state of the art in sound source extraction can be classified into three approaches:

1. Blind source separation (BSS): this approach allows for estimating a number of individual sound source components from a number of observed mixtures, by exploiting the statistical independence of the individual sources. Traditional BSS methods rely on the assumption that the number of sources is less than or equal to the number of observed mixtures, which implies that a potentially large number of microphones is required. Underdetermined BSS methods are capable of bypassing this condition, but they rely on a significant amount of prior knowledge regarding the individual sound sources. Since BSS methods tend to be computationally intensive, these are often not suited for real-time applications.
2. Computational auditory scene analysis (CASA): the aim of CASA is to analyze a sound scene in a way that mimics the human auditory system, by identifying and grouping perceptual attributes from the observed mixtures. Since CASA operates on two (binaural) microphone recordings, it is essentially an underdetermined BSS method as soon as the sound scene comprises more than two sources. While CASA has attracted the interest of many researchers, it is still considered not sufficiently mature to be used in real-life applications. Moreover, its computational requirements are typically very high.
3. Beamforming: this approach relies on the application of spatially selective filtering operations to two or more observed mixtures. There is no hard constraint on the number of observations required to separate a given number of sound sources, and moreover most beamforming implementations are computationally less demanding than the BSS or CASA approaches. However, beamforming either relies on prior knowledge about the sound source positions (in which case fixed beamformers can be applied) or requires a significant amount of additional processing for "supervision" (in the case of adaptive beamformers).

[0003] According to an aspect of the present invention there is provided an audio-processing device comprising:

an audio input, for receiving one or more audio signals detected at respective microphones, each of the audio signals comprising a mixture of a plurality of components, each component corresponding to a sound source;
a control input, for receiving, for each sound source, a desired gain factor associated with the source, by which it is desired to amplify the corresponding component;
an auxiliary signal generator, adapted to generate at least one auxiliary signal from the one or more audio signals, the at least one auxiliary signal comprising a different mixture of the components as compared with a reference one of the one or more audio signals;
a scaling coefficient calculator, adapted to calculate a set of scaling coefficients in dependence upon the desired gain factors and upon parameters of the different mixture, each scaling coefficient associated with one of the at least one auxiliary signal and optionally the reference audio signal; and
an audio synthesis unit, adapted to synthesize an output audio signal by applying the scaling coefficients to the at least one auxiliary signal and optionally the reference audio signal and to combine the results,

wherein the scaling coefficients are calculated from the desired gain factors and the parameters of the different mixture such that the synthesized output signal provides the desired gain factor for each component.

[0004] A device according to an embodiment of the invention addresses the problem of sound scene manipulation from a fundamentally different perspective, in that it allows any specified level change to be performed for each of the individual sound source components in the observed mixture(s), without relying on an explicit sound source separation. The disadvantages overcome by the device as compared to the state of the art can be explained by considering each of the three approaches highlighted already above:

1. Advantages with respect to the BSS approach: similarly to the traditional BSS approach, the processing method implemented by the device requires as many different mixtures as the number of sound source levels that it is

desired to alter independently. However, these mixtures can be generated from a smaller number of microphone recordings. For example, auxiliary mixtures can be generated by combining one microphone recording with one or more other microphone recordings. As a consequence, the method can also be used in scenarios with fewer microphones than sound sources, without a significant increase in computational burden. The proposed method has moderate computational complexity, which increases only linearly with the number of observed microphone signal samples. It is thus particularly suited for real-time applications. Finally, the method does not rely on any prior knowledge about the statistics of the individual sound sources.

2. Advantages with respect to the CASA approach: whereas the CASA approach operates on a collection of auditory features of the sound sources, the present processing method operates directly on the observed microphone signals and on a number of auxiliary signals derived from these microphone signals. Consequently, the present method does not require the estimation and detection of auditory features, which is advantageous both in terms of robustness and in terms of computational complexity.

3. Advantages with respect to the beamforming approach: whereas the beamforming approach operates only on the observed microphone signals, the present method operates on a number of auxiliary signals in addition to the microphone signals. These auxiliary signals may be generated by combining the observed microphone signals. However, there is no restriction on the mapping from the observed microphone signals to the auxiliary signals, and hence the proposed method is much more flexible than the beamforming approach. As indicated below, one embodiment of the invention may include fixed as well as adaptive beamformers for generating the auxiliary signals from the microphone signals.

[0005] One application of a method or device according to an embodiment is the enhancement of acoustic signals like speech or music. In this case, the sound scene consists of desired as well as undesired sound sources, and the aim of the sound scene manipulation comprises reducing the level of the undesired sound sources relative to the level of the desired sound sources.

[0006] According to another aspect of the invention, there is provided a handheld personal electronic device comprising a plurality of microphones; and the audio processing device referred to above.

[0007] The invention is particularly suited to mobile, handheld applications, since it has relatively light computational demands. It may therefore be usable with a mobile device having limited processing resources or may enable power consumption to be reduced.

[0008] The mobile or handheld device preferably incorporates a video recording apparatus with a visual zoom capability, and the audio processing device is preferably adapted to modify the desired gain factors in accordance with a configuration of the visual zoom. This enables the device to implement an acoustic zoom function.

[0009] The microphones are preferably omni-directional microphones.

[0010] The present device may be particularly beneficial in these circumstances, because the source separation problem is inherently more difficult when using omni-directional microphones. If the microphones are unidirectional, there will often be significant selectivity (in terms of signal power) between the sources among the diverse audio signals. This can make the manipulation task easier. The present device is able to work also with omnidirectional microphones, where there will be less selectivity in the raw audio signals. The present device is therefore more flexible. For example, it can exploit spatial selectivity by means of beamforming techniques, but it is not limited to spatial selectivity through the use of unidirectional microphones.

[0011] According to a further aspect of the invention, there is provided a method of processing audio signals comprising:

receiving one or more audio signals detected at respective microphones, each of the audio signals comprising a mixture of a plurality of components, each component corresponding to a sound source;

receiving, for each sound source, a desired gain factor associated with the source, by which it is desired to amplify the corresponding component;

generating at least one auxiliary signal from the one or more audio signals, the at least one auxiliary signal comprising a different mixture of the components as compared with a reference one of the one or more audio signals;

calculating a set of scaling coefficients in dependence upon the desired gain factors and upon parameters of the different mixture, each scaling coefficient associated with one of the at least one auxiliary signal and optionally the reference audio signal; and

synthesizing an output audio signal by applying the scaling coefficients to the at least one auxiliary signal and optionally the reference audio signal and combining the results,

wherein the scaling coefficients are calculated from the desired gain factors and the parameters of the different mixture such that the synthesized output signal provides the desired gain factor for each component.

[0012] The parameters of the different mixture may be reweighting factors, which relate the levels of the components in the at least one auxiliary signal to their respective levels in the reference audio signal.

[0013] The method is particularly relevant to configurations with more than one microphone. Sound from all of the sound sources is detected at each microphone. Therefore, each sound source gives rise to a corresponding component in each audio signal. The number of sources may be less than, equal to, or greater than the number of audio signals (which is equal to the number of microphones). The sum of the number of audio signals and the number of auxiliary signals should be at least equal to the number of sources which it is desired to independently control.

[0014] Each auxiliary signal contains a different mixture of the components. That is, the components occur with different amplitude in each of the auxiliary signals (according to the reweighting factors). In other words, the auxiliary signals and the audio signals should be linearly independent; and the sets of reweighting factors which relate the signal components to each auxiliary signal should also be linearly independent of one another.

[0015] Explicit source separation is not necessary. Preferably the levels of the source signal components in the auxiliary signals are varied by a power ratio in the range -40dB to +6dB, more preferably -30dB to 0dB, still more preferably -25dB to 0dB, compared to their levels in the reference audio signal(s).

[0016] In the step of synthesizing the output signal, a scaling coefficient is preferably applied to the reference original audio signal and the result is combined with the scaled auxiliary signals.

[0017] The scaled auxiliary signals and/or scaled audio signals may be combined by summing them.

[0018] In general, in practice, the scaling coefficients and the desired gain factors will have different values (and may be different in number). They would only be identical if the auxiliary signals were to achieve perfect separation of the sources, which is usually impossible in practice. Each desired gain factor corresponds to the desired volume (amplitude) of a respective one of the sound-sources. On the other hand, the scaling coefficients correspond to the auxiliary signals and/or input audio signals. The number of reweighting factors is equal to the product of the number of signal components and the number of auxiliary signals, since in general each auxiliary signal will comprise a mixture of all of the signal components.

[0019] Preferably, the desired gain factors; the reweighting factors and the scaling coefficients are related by a linear system of equations; and the step of calculating the set of scaling coefficients comprises solving the system of equations.

[0020] For example, the step of calculating the set of scaling coefficients may comprise: calculating the inverse of a matrix of the reweighting factors; and multiplying the desired gain factors by the result of this inversion calculation.

[0021] The reweighting factors may be formed into a matrix and the inverse of this matrix may be calculated explicitly. Alternatively, the inverse may be calculated implicitly by equivalent linear algebraic calculations. The result of the inversion may be expressed as a matrix, though this is not essential.

[0022] The at least one auxiliary signal may be a linear combination of any of: one or more of the audio signals; one or more shifted versions of the audio signals; and one or more filtered versions of the audio signals.

[0023] The at least one auxiliary signal may be generated by at least one of: fixed beamforming; adaptive beamforming; and adaptive spectral modification.

[0024] Here, fixed beamforming means a spatially selective signal processing operation with a time-invariant spatial response. Adaptive beamforming means a spatially selective signal processing operation with a time-varying spatial response. Adaptive spectral modification means a frequency-selective signal processing operation with a time-varying frequency response, such as the class of methods known in the art as adaptive spectral attenuation or adaptive spectral subtraction. An adaptive spectral modification process typically does not exploit spatial diversity, but only frequency diversity among the signal components.

[0025] These are advantageous examples of ways to create the auxiliary signals. Fixed beamforming may be beneficial when there is some prior expectation that one or more of the sound sources is localised and located in a predetermined direction relative to a set of microphones. The fixed beamformer will then modify the power of the corresponding signal component, relative to other components.

[0026] Adaptive beamforming may be beneficial when a localised sound source is expected, but its orientation relative to the microphone(s) is unknown.

[0027] Adaptive spectral modification (for example, by attenuation) may be useful when sound sources can be discriminated to some extent by their spectral characteristics. This may be the case for a diffuse noise source, for example.

[0028] The methods of generating the auxiliary signal or signals are preferably chosen according to the expected sound environment in a given application. For example, if several sources in known directions are expected, it may be appropriate to use multiple fixed beamformers. If multiple moving sources are expected, multiple adaptive beamformers may be beneficial. In this way - as will be apparent to those skilled in the art - one or more instances of different means of generating the auxiliary signal may be combined, in embodiments.

[0029] Optionally, a first auxiliary signal is generated by a first method; a second auxiliary signal is generated by a second, different method; and the second auxiliary signal is generated based on an output of the first method.

[0030] For example, the fixed beamforming may be adapted to emphasize sounds originating directly in front of the microphone or microphone array. For example, this may be useful when the microphone is used in conjunction with a camera, because the camera (and therefore the microphone) is likely to be aimed at a subject who is one of the sound sources.

[0031] An output of the fixed beamformer may be input to the adaptive beamformer. This may be a noise reference output of the fixed beamformer, wherein the power ratio of a component originating from the fixed direction is reduced relative to other components. It is advantageous to use this signal in the adaptive beamformer, in order to find a (remaining) localised source in an unknown direction, because the burden on the adaptive beamformer to suppress the fixed signals may be reduced.

[0032] An output of the adaptive beamformer may be input to the adaptive spectral modification.

[0033] Typically, neither of the beamformers nor an adaptive spectral attenuator will be sufficiently selective to separate individual sources from the mixture. In this context, the method of the invention may be seen as a flexible framework for combining weak separators, to allow an arbitrary desired weighting on sound sources. The individual operations of beamforming or spectral modification preferably cause a change in the signal power of individual sound source components in the range -25dB to 0dB. This refers to the input/output power ratio of each operation, ignoring cascade effects due to the output of one unit being connected to the input of another

[0034] The method may optionally comprising: synthesizing a first output audio signal by applying scaling coefficients to a first reference audio signal and at least one first auxiliary signal and combining the results; and synthesizing a second output audio signal by applying scaling coefficients to a second, different reference audio signal and at least one second auxiliary signal and combining the results.

[0035] This may be particularly useful for generating binaural (for example, stereo) outputs. The at least one first auxiliary signal and at least one second auxiliary signal may be the same or different signals. The two different reference audio signals should be selected from appropriately arranged microphones, for a desired stereo effect.

[0036] In a similar way, the method can be extended to synthesize an arbitrarily greater number of outputs, as desired for any particular application.

[0037] The sound sources may comprise one or more localised sound sources and a diffuse noise field.

[0038] The desired gain factors may be time-varying.

[0039] The method is particularly well suited to real-time implementation, which means that the desired gain can be adjusted dynamically. This may be useful for example for dynamically balancing changing sound sources, or for acoustic zooming.

[0040] In a sound scene consisting of multiple desired sound sources, one often encounters the problem that the levels of the different sources are not sufficiently balanced in the microphone recordings - for example, if one of the sources is positioned closer to the microphone array than the others. In a static scenario, the sound scene can be balanced using time-invariant gain factors, while in a dynamic scenario (that is, with moving or temporally modulated sound sources) the use of time-varying gain factors is more relevant.

[0041] The desired gain factors can be chosen in dependence upon the state of a visual zoom function.

[0042] In applications where joint audio and video recordings are made (for example, camcorder or video-phone applications), it may be beneficial to match the auditory and visual cues in the recordings to obtain an easier and/or faster multisensory integration. A key example is the process of manipulating the sound scene such that it properly matches the video zooming operations. For example, when zooming in on a particular subject, the sound level of this subject should increase accordingly while keeping the level of the other sound sources constant. In this case, the desired gain factor corresponding to the sound source in front of the camera will be increased over time, while the other gain factors are time-invariant.

[0043] Also provided is a computer program comprising computer program code means adapted to perform all the steps of a method as described above, when said program is run on a computer; and such a computer program embodied on a computer readable medium.

[0044] The invention will now be described by way of example with reference to the accompanying drawings, in which:

Fig. 1 shows a block diagram of an audio processing device according to an embodiment;

Fig. 2 shows in greater detail an auxiliary signal generator and audio synthesis unit suitable for a monaural implementation of the embodiment of Fig. 1;

Fig. 3 shows in greater detail an auxiliary signal generator and audio synthesis unit suitable for a binaural (stereo) implementation of the embodiment of Fig. 1; and

Fig. 4 is a flowchart of a method according to an embodiment.

[0045] In the following, a theoretical explanation of a method according to an embodiment will first be given, along with an indication of the conditions under which this theory can be used for sound scene manipulation.

[0046] Consider a sound scene consisting of M localized sound sources $s_m(t)$, $m = 1,...,M$ positioned in different directions in the three-dimensional plane (as characterized by the azimuth-elevation angle pairs (θ_m,ϕ_m), $m = 1,...,M$), in addition to a diffuse sound field that cannot be attributed to a single sound source or direction. Further to this, consider a microphone array consisting of N microphones ($N \geq 2$) and having an arbitrary three-dimensional geometry. Each of the microphones may have a different frequency- and angle-dependent response, as defined by

$$A_n(\omega,\theta,\phi) = a_n(\omega,\theta,\phi)e^{-j\psi_n(\omega,\theta,\phi)}, \quad n = 0,\ldots,N-1. \qquad (1)$$

[0047] The acoustic response (including the effect of the direct path time delay as well as reverberation) of a sound source at angle (θ,ϕ) to each of the microphones is given by

$$F_n(\omega,\theta,\phi) = f_n(\omega,\theta,\phi)e^{-j\xi_n(\omega,\theta,\phi)}, \quad n = 0,\ldots,N-1. \qquad (2)$$

[0048] For ease of notation, we introduce the joint acoustic and microphone response, defined as

$$G_n(\omega,\theta,\phi) = A_n(\omega,\theta,\phi)F_n(\omega,\theta,\phi), \quad n = 0,\ldots,N-1. \qquad (3)$$

[0049] Using the above definitions, we can express each of the *N* audio signals $U_n(\omega)$ detected at the microphones as a function of the localized sound sources and the diffuse sound field in the frequency domain as follows:

$$U_n(\omega) = U_n^{(0)}(\omega) + \sum_{m=1}^{M} \underbrace{G_n(\omega,\theta_m,\phi_m)S_m(\omega)}_{U_n^{(m)}(\omega)}, \quad n = 0,\ldots,N-1 \qquad (4)$$

where $U_n^{(0)}(\omega)$ denotes the diffuse noise component. The above relation can equivalently be written in the time domain as follows,

$$u_n(t) = u_n^{(0)}(t) + \sum_{m=1}^{M} u_n^{(m)}(t). \qquad (5)$$

[0050] The aim of the envisaged sound scene manipulation is to produce N manipulated signals, or audio output signals, $\zeta_n(t)$, in which each of the levels of the individual sound source components is changed in a user-specified way as compared to the respective levels in the nth microphone signal. Mathematically, the aim is to produce the signals

$$\zeta_n(t) = g_n^{(0)}(t)u_n^{(0)}(t) + \sum_{m=1}^{M} g_n^{(m)}(t)u_n^{(m)}(t), \quad n = 0,\ldots,N-1 \qquad (6)$$

where $g_n^{(m)}(t)$, $m = 0,\ldots,M$ denote the user-specified time-varying gain factors for the different sound source components. Hereinafter, these will be referred to as the "desired gain factors".

[0051] Suppose that one could generate M auxiliary signals $x_n^{(p)}(t),\ p = 1,\ldots,M$, in which the different sound source components have been arbitrarily reweighted with respect to the corresponding components in the microphone signal $u_n(t)$, that is,

$$x_n^{(p)}(t) = \sum_{m=0}^{M} \gamma_n^{(p,m)} u_n^{(m)}(t) \qquad (7)$$

[0052] Here, each of the reweighting factors is by definition equal to the square root of the power ratio of the corresponding sound source components, that is,

$$\gamma_n^{(p,m)} = \frac{\sigma_{x_n^{(p)}}}{\sigma_{u_n^{(m)}}} = \sqrt{\frac{E\left\{\left(x_n^{(p)}\right)^2\right\}}{E\left\{\left(u_n^{(m)}\right)^2\right\}}}. \qquad (8)$$

[0053] The *n*th manipulated signal (output audio signal) can now be calculated as a weighted sum of the nth microphone signal and the auxiliary signals $x_n^{(p)}(t),\ p = 1,\ldots,M$ defined above, that is,

$$\zeta_n(t) = a_n^{(0)}(t) u_n(t) + \sum_{p=1}^{M} a_n^{(p)}(t) x_n^{(p)}(t). \qquad (9)$$

[0054] By using the relations in equations (5) and (7), the expression for the calculated *n*th manipulated signal in equation (9) can be shown to be equivalent to the expression for the desired nth manipulated signal in equation (6) if the weights $a_n^{(p)}(t),\ p = 0,\ldots,M$ satisfy the following relationship,

$$\begin{bmatrix} a_n^{(0)}(t) \\ a_n^{(1)}(t) \\ \vdots \\ a_n^{(M)}(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \gamma_n^{(1,0)} & \ldots & \gamma_n^{(M,0)} \\ 1 & \gamma_n^{(1,1)} & \ldots & \gamma_n^{(M,1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \gamma_n^{(1,M)} & \ldots & \gamma_n^{(M,M)} \end{bmatrix}^{-1}}_{\Gamma^{-1}} \begin{bmatrix} g_n^{(0)}(t) \\ g_n^{(1)}(t) \\ \vdots \\ g_n^{(M)}(t) \end{bmatrix} \qquad (10)$$

[0055] This implies that a unique set of weight trajectories $a_n^{(p)}(t),\ p = 0,\ldots,M,\ \forall t$ can be calculated that *exactly* produces the desired sound scene manipulation. Herein, the weight trajectories $a_n^{(p)}(t),\ p = 0,\ldots,M,\ \forall t$ are also referred to as "scaling coefficients".

[0056] There are two conditions for exact reproduction of the effect of an arbitrary set of desired gain factors $g_n^{(m)}(t)$, according to equation (10):

1. The reweighting matrix r should be of full rank,
2. The reweighting factors $\gamma_n^{(p,m)}$ should be known.

[0057] Loosely speaking, the first condition requires that the microphone signal $u_n(t)$ and the auxiliary signals $x_n^{(p)}(t),\ p = 1,\ldots,M$ should be linearly independent (which leads to linear independent columns in Γ) and requires that the reweighting of the different sound source components in each of the auxiliary signals $x_n^{(p)}(t),\ p = 1,\ldots,M$ should be linearly independent (which leads to linear independent rows in r). The reweighting factors can be calculated

or estimated depending on the embodiment of the invention, as described in greater detail below.

[0058] Note that equation (7) above is a model for the auxiliary signals which will usually be satisfied only approximately, in practice. In the embodiments described below, the auxiliary signals will be derived from the various microphone signals. Therefore, they will be composed of filtered versions of the sound source components, instead of the unfiltered ("dry") sound source components themselves suggested by equation (7).

[0059] If the model of equation (7) could be satisfied precisely, exact recovery of a single sound source component would be possible (by choosing the desired gain factors appropriately). In the embodiment to be described below, this would demand the design of ideal beamformers that have a flat frequency response within the bandwidth of the source component of interest, and demand that the diffuse noise has no spectral overlap with the source component of interest. In practice, these restrictions are usually not met, and as a consequence the auxiliary signals will be linear combinations of filtered versions of the original sound source components (with non-uniform frequency response), rather than linear combinations of the original sound source components. This makes the exact recovery of a single sound source component impossible; however, this is a shortcoming of the practical embodiment rather than the theoretical method.

[0060] In the following, without loss of generality, an exemplary scenario will be considered in which the sound field in the acoustic environment is assumed to consist of four contributions coming from a different azimuthal directions:

1) a front sound source $s_F(t)$, which is considered to be the desired sound source and is located in front of the camera at an angle $\theta_F = 0$ (by definition);
2) a back sound source $s_B(t)$, which may or may not be a desired sound source, corresponding to the sound produced by the camera operator (if any) at an angle $\theta_B = 180$ degrees;
3) a number of localized interfering sound sources $\{s_I^{(i)}(t)\}_{i=1}^{N_I}$ which are considered to be undesired and originate from (unknown) directions $\theta_I^{(i)}$ different from the front and back directions; and
4) a diffuse noise field, which cannot be attributed to a single sound source or direction, and which is also considered to be undesired.

[0061] The number of localized interfering sound sources is taken to be one, for the purposes of this explanation. Furthermore, in this example, it is assumed that the capture device is equipped with two or more microphones. Those skilled in the art will appreciate that none of these assumptions should be taken to limit the scope of the invention.

[0062] If the nth microphone signal $u_n(t)$ is decomposed in the time domain as:

$$u_n(t) = u_n^{(F)}(t) + u_n^{(B)}(t) + u_n^{(I)}(t) + u_n^{(N)}(t)$$

then the corresponding desired output of the algorithm can be written as follows:

$$\zeta_n(t) = g_F(t)u_n^{(F)}(t) + g_B(t)u_n^{(B)}(t) + g_I(t)u_n^{(I)}(t) + g_N(t)u_n^{(N)}(t)$$

where $g_F(t)$, $g_B(t)$, $g_I(t)$, and $g_N(t)$ denote the desired gain factors for the different sound source components. Note that one is not necessarily interested in calculating N output signals of the algorithm. Typically, the focus is on obtaining a mono or stereo output, which implies that the relation above only needs to be considered for one or two particular values of n, say n_1 (and n_2).

[0063] Nevertheless, all N microphone signals will typically be used to obtain an estimate of the two output signals, $\zeta_{n1}(t), \zeta_{n2}(t)$. Also note that we have not included the output signal index n in the notation of the gain factors in the equation above, since typically the same gain factors will be used for the different output signals of the algorithm. (Of course, this is not essential).

[0064] Conventionally, it would be expected that the algorithm needs to perform some kind of source separation to isolate the different sound source components. However, since we are not interested in the separated sound source components, but rather in a mixture in which the levels of these components have been adjusted as compared to the microphone signals, an explicit source separation is not required. Let us denote three auxiliary signals as $x_n(t)$, $y_n(t)$, and $z_n(t)$, in which the different sound source components have been arbitrarily reweighted (by reweighting factors γ) with respect to the corresponding components in the microphone signal $u_n(t)$, that is:

$$x_n(t) = \gamma^{(F)}_{x_n,u_n} u^{(F)}_n(t) + \gamma^{(B)}_{x_n,u_n} u^{(B)}_n(t) + \gamma^{(I)}_{x_n,u_n} u^{(I)}_n(t) + \gamma^{(N)}_{x_n,u_n} u^{(N)}_n(t)$$

$$y_n(t) = \gamma^{(F)}_{y_n,u_n} u^{(F)}_n(t) + \gamma^{(B)}_{y_n,u_n} u^{(B)}_n(t) + \gamma^{(I)}_{y_n,u_n} u^{(I)}_n(t) + \gamma^{(N)}_{y_n,u_n} u^{(N)}_n(t)$$

$$z_n(t) = \gamma^{(F)}_{z_n,u_n} u^{(F)}_n(t) + \gamma^{(B)}_{z_n,u_n} u^{(B)}_n(t) + \gamma^{(I)}_{z_n,u_n} u^{(I)}_n(t) + \gamma^{(N)}_{z_n,u_n} u^{(N)}_n(t).$$

[0065] The output signal of the algorithm can now be calculated as a linear combination of the nth microphone signal and the auxiliary signals $x_n(t)$, $y_n(t)$, and $Z_n(t)$ defined above, that is:

$$\zeta_n(t) = a^{(0)}_n(t) u_n(t) + a^{(1)}_n(t) x_n(t) + a^{(2)}_n(t) y_n(t) + a^{(3)}_n(t) z_n(t).$$

[0066] This corresponds to equation (9) above. The corresponding form of equation (10) is:

$$\begin{bmatrix} a^{(0)}_n(t) \\ a^{(1)}_n(t) \\ a^{(2)}_n(t) \\ a^{(3)}_n(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \gamma^{(F)}_{x_n,u_n} & \gamma^{(F)}_{y_n,u_n} & \gamma^{(F)}_{z_n,u_n} \\ 1 & \gamma^{(B)}_{x_n,u_n} & \gamma^{(B)}_{y_n,u_n} & \gamma^{(B)}_{z_n,u_n} \\ 1 & \gamma^{(I)}_{x_n,u_n} & \gamma^{(I)}_{y_n,u_n} & \gamma^{(I)}_{z_n,u_n} \\ 1 & \gamma^{(N)}_{x_n,u_n} & \gamma^{(N)}_{y_n,u_n} & \gamma^{(N)}_{z_n,u_n} \end{bmatrix}^{-1}}_{\Gamma^{-1}} \begin{bmatrix} g_F(t) \\ g_B(t) \\ g_I(t) \\ g_N(t) \end{bmatrix}$$

[0067] This enables the scaling factors, a, to be calculated, provided the re-weighting factors are known. The estimation of the re-weighting factors will be described in greater detail below. Before that, two embodiments of the invention will be described.

[0068] Both embodiments have the general structure shown in the block diagram of Fig. 1. An array of microphones 4 produces a corresponding plurality of audio signals 6. These are fed as input to an auxiliary signal generator 10. The auxiliary signal generator generates auxiliary signals, each comprising a mixture of the same sound source components detected by the microphones 4, but with the components present in the mixture with different relative strengths (as compared with their levels in the original audio signals 6). In the embodiments described below, these auxiliary signals are derived by processing combinations of the audio signals 6 in various ways. The auxiliary signals and the input audio signals 6 are fed as inputs to an audio synthesis unit 20. This unit 20 applies scaling coefficients to the signals and sums them, to produce output signals 40. In the output signals 40, the sound source components are present with desired strengths. These desired strengths are expressed by gain factors 8, which are input to a scaling coefficient calculator 30. The scaling coefficient calculator 30 converts the desired gains {g(t)} into a set of scaling coefficients {a(t)}. Each of the desired gains is associated with a sound source detectable at the microphones 4; whereas each of the scaling coefficients is associated with one of the auxiliary signals. The scaling coefficient calculator 30 exploits knowledge about the parameters of the auxiliary signals to transform from desired gains {g(t)} to suitable scaling coefficients {a(t)}.

[0069] In the first embodiment the goal is to obtain a monaural (mono) output signal. Fig. 2 shows a block structure for the calculation of the auxiliary signals $x_n(t)$, $y_n(t)$, and $z_n(t)$ required in the algorithm.

[0070] In Fig. 2, the auxiliary signal generator 10 consists of three functional blocks 210, 212, 214:

1) *Fixed beamformer 210:* the purpose of this block is to perform reweighting of the sound source components of which the source direction is known a priori - that is, the front and back sound sources. The power ratios of these components are altered by the fixed beamformer, both relative to each other and relative to the other sound source components.

2) Adaptive beamformer 212: this block serves to perform reweighting of the localized interfering sound source(s). This necessarily requires an adaptive beamforming algorithm since the interfering sound source direction is unknown.

3) *Adaptive spectral attenuation 214:* this block reweights the diffuse noise field, by exploiting its assumed spectral diversity with reference to the localized sound source components.

[0071] The audio synthesis unit 20 is indicated by the dashed box 220. This produces the output signal $\zeta_0(t)$ as a weighted summation of the auxiliary signals x_0, y_0, and z_0, as well as the reference audio signal u_0. The weights are the scaling coefficients, a, derived by the scaling coefficient calculator 30 (not shown in Fig. 2).

[0072] Note that in the mono output case of Fig. 2, some of the auxiliary signals (more particularly $x_n(t)$ and $y_n(t)$ for $n > 0$) are not explicitly used to calculate the output signal. However, these signals are used internally in the adaptive beamformer and adaptive spectral attenuation algorithms. More particularly, the signals $x_n(t)$, $n > 0$ at the output of the fixed beamformer will be constructed to be "noise reference signals"; that is, signals in which the desired (front and optionally back) sound sources have been suppressed and which are used subsequently in the adaptive beamformer to estimate the localized interfering sound source component in the primary output signal $x_0(t)$ of the fixed beamformer. The signal $y_1(t)$ is then constructed to be a "diffuse noise reference" that is used by the adaptive spectral attenuation algorithm to estimate the diffuse noise component in the primary output signal $y_0(t)$ of the fixed beamformer.

[0073] Because of the above discrimination between the primary beamformer output signals $x_0(t)$ and $y_0(t)$, on the one hand; and the other beamformer output signals $x_n(t)$ and $y_n(t)$ with $n > 0$, on the other hand, a stereo output signal should preferably not be created by calculating $\zeta_0(t)$ and $\zeta_1(t)$ using these auxiliary signals.

[0074] Instead, in the second embodiment, the block structure shown in Fig. 3 is used for the stereo case. Here, the stereo output signals are calculated as follows:

$$\zeta_0(t) = a_0^{(0)}(t)u_0(t) + a_0^{(1)}(t)x_0(t) + a_0^{(2)}(t)y_0(t) + a_0^{(3)}(t)z_0(t)$$

$$\zeta_1(t) = a_1^{(0)}(t)u_1(t) + a_1^{(1)}(t)x_0(t) + a_1^{(2)}(t)y_0(t) + a_1^{(3)}(t)z_0(t)$$

[0075] That is, the same set of auxiliary signals is used for generating both stereo outputs, but a different reference audio signal, $u_n(t)$, is used in each case. This computation is performed by the audio synthesis unit 320 indicated by the dashed box.

[0076] In the case that $N > 2$ (that is, when the array consists of more than two microphones), one should select $u_0(t)$ and $u_1(t)$ to be those two microphone signals that are best suited to deliver a stereo image. As will be apparent to those skilled in the art, this will typically depend on the placement of the microphones.

[0077] Note that, due to the particular structure shown in Fig. 5, the weight calculation for the second output signal $\zeta_1(t)$ should be slightly altered, to:

$$\begin{bmatrix} a_1^{(0)}(t) \\ a_1^{(1)}(t) \\ a_1^{(2)}(t) \\ a_1^{(3)}(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \gamma_{x_0,u_1}^{(F)} & \gamma_{y_0,u_1}^{(F)} & \gamma_{z_0,u_1}^{(F)} \\ 1 & \gamma_{x_0,u_1}^{(B)} & \gamma_{y_0,u_1}^{(B)} & \gamma_{z_0,u_1}^{(B)} \\ 1 & \gamma_{x_0,u_1}^{(I)} & \gamma_{y_0,u_1}^{(I)} & \gamma_{z_0,u_1}^{(I)} \\ 1 & \gamma_{x_0,u_1}^{(N)} & \gamma_{y_0,u_1}^{(N)} & \gamma_{z_0,u_1}^{(N)} \end{bmatrix}^{-1}}_{\Gamma^{-1}} \begin{bmatrix} g_F(t) \\ g_B(t) \\ g_I(t) \\ g_N(t) \end{bmatrix}$$

[0078] Meanwhile, the weights for the primary output signal $\zeta_1(t)$ can be calculated as before, with $n = 0$.

[0079] As the equations above show, the scaling coefficient calculator 30 uses knowledge of the reweighting factors $\gamma_n^{(p,m)}$ to derive the scaling coefficients, a(t), from the desired gains, g(t). In the presently described embodiments, the reweighting factors are found by using knowledge of the characteristics of the various blocks 210, 212, 214 in the auxiliary signal generator. Preferably, the reweighting factors are determined offline.

[0080] Examples of the calculation of the reweighting factors will be described below. These examples rely on a frequency-domain characterisation of the auxiliary signal generator blocks 210, 212, 214.

[0081] The input-output relation of the three functional blocks in the block structure can be described in the frequency domain as follows. The fixed beamformer can be specified by an N xN transfer function matrix $\mathbf{W}_1(\omega)$, that is,

$$\mathbf{X}(\omega) = \mathbf{W}_1^H(\omega)\mathbf{U}(\omega)$$

where

$$\mathbf{X}(\omega) = \begin{bmatrix} X_0(\omega) & \dots & X_{N-1}(\omega) \end{bmatrix}^T$$

$$\mathbf{W}_1(\omega) = \begin{bmatrix} W_{1,(1,1)}(\omega) & \dots & W_{1,(1,N)}(\omega) \\ \vdots & \ddots & \vdots \\ W_{1,(N,1)}(\omega) & \dots & W_{1,(N,N)}(\omega) \end{bmatrix}$$

and **U**(ω) is defined as

$$\mathbf{U}(\omega) = \begin{bmatrix} U_0(\omega) & \dots & U_{N-1}(\omega) \end{bmatrix}^T$$

[0082] The adaptive beamformer can be specified by an N x 1 transfer function vector $\mathbf{W}_2(\omega)$ that defines the relation between the adaptive beamformer input and its primary output signal:

$$Y_0(\omega) = \mathbf{W}_2^H(\omega)\mathbf{X}(\omega)$$

where

$$\mathbf{W}_2(\omega) = \begin{bmatrix} W_{2,(1)}(\omega) & \dots & W_{2,(N)}(\omega) \end{bmatrix}^T$$

[0083] As explained earlier, the secondary adaptive beamformer output signal should ideally be an estimate of the diffuse noise component in the primary adaptive beamformer output signal. The most straightforward approach is to choose the secondary output signal to be equal to one of the noise references at the output of the fixed beamformer- for example, $Y_1(\omega) = X_1(\omega)$. Alternatively, one could attempt to remove the localized interfering sound source component from the secondary adaptive beamformer output signal, however, this approach is not used in the present embodiments. The adaptive spectral attenuation can finally be specified using a scalar transfer function $W_3(\omega)$,that is,

$$Z_0(\omega) = W_3(\omega)Y_0(\omega)$$

[0084] Using the above input-output relations, we can derive expressions for the different localized sound source components in the primary auxiliary signals $X_0(\omega)$, Y0(ω), and $Z_0(\omega)$ as a function of the corresponding dry sound source signals $S_F(\omega)$, $S_B(\omega)$, and $S_I(\omega)$,

$$X_0^{(c)}(\omega) = \mathbf{W}_{1,(:,1)}^H(\omega)\mathbf{G}(\omega,\theta_c)S_c(\omega)$$

$$Y_0^{(c)}(\omega) = \mathbf{W}_2^H(\omega)\mathbf{W}_1^H(\omega)\mathbf{G}(\omega,\theta_c)S_c(\omega)$$

$$Z_0^{(c)}(\omega) = W_3(\omega)\mathbf{W}_2^H(\omega)\mathbf{W}_1^H(\omega)\mathbf{G}(\omega,\theta_c)S_c(\omega)$$

[0085] where c represents the component F, B, or I, and $W_{1,(:,1)}(\omega)$ denotes the first column of $W_1(\omega)$. Similarly, the diffuse noise component in the primary auxiliary signals can be expressed as a function of the diffuse noise components

in the microphone signals,

$$X_0^{(N)}(\omega) = \mathbf{W}_{1,(:,1)}^H(\omega)\mathbf{U}^{(N)}(\omega)$$
$$Y_0^{(N)}(\omega) = \mathbf{W}_2^H(\omega)\mathbf{W}_1^H(\omega)\mathbf{U}^{(N)}(\omega)$$
$$Z_0^{(N)}(\omega) = W_3(\omega)\mathbf{W}_2^H(\omega)\mathbf{W}_1^H(\omega)\mathbf{U}^{(N)}(\omega)$$

[0086] We will now make the following assumptions, to simplify the calculation of the reweighting factors:

1) the joint acoustic and microphone responses have a flat magnitude response within the bandwidth and in the direction of the different sound source components, i.e.,

$$\forall\omega : S_c(\omega) \neq 0, U_n^{(N)}(\omega) \neq 0 \Rightarrow |G_n(\omega,\theta_c)| \equiv |G_n(\theta_c)|,$$
$$n = 0,\ldots,N-1,\ c = F,B,I$$

2) the fixed and adaptive beamformers have a flat magnitude response within the bandwidth and in the direction of the different sound source components, that is,

$$\forall\omega : S_c(\omega) \neq 0, U_n^{(N)}(\omega) \neq 0 \Rightarrow \begin{cases} |W_{1,(m,n)}(\omega)| & \equiv & |W_{1,(m,n)}|, \\ |W_{2,(n)}(\omega)| & \equiv & |W_{2,(n)}|, \end{cases}$$
$$m = 1,\ldots,N,\ n = 1,\ldots,N,\ c = F,B,I$$

3) the diffuse noise spectrum does not overlap with the spectra of the different localized sound sources,

$$\nexists\omega : S_c(\omega) \neq 0 \ \&\ U_n^{(N)}(\omega) \neq 0, \quad n = 0,\ldots,N-1,\ c = F,B,I$$

4) the adaptive spectral attenuation magnitude response is flat within the bandwidth of the localized sound sources and within the bandwidth of the diffuse noise,

$$\forall\omega : S_c(\omega) \neq 0 \Rightarrow |W_3(\omega)| \equiv |W_3^{(c)}|, \quad c = F,B,I$$
$$\forall\omega : U_n^{(N)}(\omega) \neq 0 \Rightarrow |W_3(\omega)| \equiv |W_3^{(N)}|, \quad n = 0,\ldots,N-1$$

5) the diffuse noise power in each of the microphone signals is equal,

$$\sigma^2_{u_0^{(N)}} = \ldots = \sigma^2_{u_{N-1}^{(N)}}$$

[0087] Under these assumptions, the signal powers of the different sound source components in the microphone and auxiliary signals can be estimated as follows:

$$\sigma^2_{u_n^{(c)}} = |G_n(\theta_c)|^2\sigma^2_{s_c}, \quad n = 0,\ldots,N-1,\ c = F,B,I$$

$$\sigma^2_{x_0^{(c)}} = |\mathbf{W}^H_{1,(:,1)}\mathbf{G}(\theta_c)|^2\sigma^2_{s_c}, \quad c = F,B,I$$

$$\sigma^2_{y_0^{(c)}} = |\mathbf{W}_2^H\mathbf{W}_1^H\mathbf{G}(\theta_c)|^2\sigma^2_{s_c}, \quad c = F,B,I$$

$$\sigma^2_{z_0^{(c)}} = |W_3^{(c)}|^2|\mathbf{W}_2^H\mathbf{W}_1^H\mathbf{G}(\theta_c)|^2\sigma^2_{s_c}, \quad c = F,B,I$$

$$\sigma^2_{x_0^{(N)}} = \|\mathbf{W}_{1,(:,1)}\|_2^2\sigma^2_{u_0^{(N)}}$$

$$\sigma^2_{y_0^{(N)}} = \|\mathbf{W}_1\mathbf{W}_2\|_2^2\sigma^2_{u_0^{(N)}}$$

$$\sigma^2_{x_0^{(N)}} = |W_3^{(N)}|^2\|\mathbf{W}_1\mathbf{W}_2\|_2^2\sigma^2_{u_0^{(N)}}$$

and consequently, the reweighting factors can be calculated as

$$\gamma^{(c)}_{x_0,u_n} = \frac{|\mathbf{W}^H_{1,(:,1)}\mathbf{G}(\theta_c)|}{|G_n(\theta_c)|}, \quad n = 0,\ldots,N-1,\ c = F,B,I$$

$$\gamma^{(c)}_{y_0,u_n} = \frac{|\mathbf{W}_2^H\mathbf{W}_1^H\mathbf{G}(\theta_c)|}{|G_n(\theta_c)|}, \quad n = 0,\ldots,N-1,\ c = F,B,I$$

$$\gamma^{(c)}_{z_0,u_n} = \frac{|W_3^{(c)}||\mathbf{W}_2^H\mathbf{W}_1^H\mathbf{G}(\theta_c)|}{|G_n(\theta_c)|}, \quad n = 0,\ldots,N-1,\ c = F,B,I$$

$$\gamma^{(N)}_{x_0,u_n} = \|\mathbf{W}_{1,(:,1)}\|_2, \quad n = 0,\ldots,N-1$$

$$\gamma^{(N)}_{y_0,u_n} = \|\mathbf{W}_1\mathbf{W}_2\|_2, \quad n = 0,\ldots,N-1$$

$$\gamma^{(N)}_{z_0,u_n} = |W_3^{(N)}|\|\mathbf{W}_1\mathbf{W}_2\|_2, \quad n = 0,\ldots,N-1$$

[0088] Finally, note that from a computational point of view, in some applications, it may be undesirable to calculate the reweighting factors online (in real-time) using the preceding formulae. A more efficient approach involves setting the values of the reweighting factors off-line (in advance), making use of the fixed beamformer response (known a priori) and of heuristics about the behaviour of the adaptive beamformer and spectral attenuation response. The values chosen can be approximations of the theoretical values predicted by the equations above. For example, the values may be set heuristically in 5dB steps. In many applications, the method will be largely insensitive to 5dB or 10dB deviations from the precise theoretical values.

[0089] The design of the fixed beamformer in an exemplary embodiment will now be described.

[0090] As explained previously above, the fixed beamformer creates a primary output signal $X_0(\omega)$ that spatially enhances the front sound source signal, as well as a number of other output signals $X_n(\omega)$, n > 0 that serve as "noise references" for the adaptive beamformer. Here, we will first discuss the design of the so-called front source beamformer (FSB), and afterwards we will explain the design of the so-called blocking matrix (BM).

[0091] Depending on the kind of spatial enhancement one wants to achieve for the front sound source, different fixed beamformer design methods could be employed for the FSB; for example, an array pattern synthesis approach, or a differential or superdirective design method. These methods themselves are known in the art. In the present embodiment, we will adopt a superdirective (SD) design method, which is recommendable when the aim is to maximize the directivity factor of the microphone array - that is, to maximize the array gain in the presence of a diffuse noise field. The frequency-domain SD design equation for the FSB can be found in S. Doclo and M. Moonen ("Superdirective beamforming robust against microphone mismatch," IEEE Trans. Audio Speech Lang. Process., vol. 15, no. 2, pp. 617-631, Feb. 2007):

$$\mathbf{W}_{1,(:,1)}(\omega) = \frac{\left(\tilde{\mathbf{\Phi}}_{\mathbf{U}}^{(N)} + \mu \mathbf{I}_N\right)^{-1} \mathbf{G}(\omega, \theta_F)}{\mathbf{G}^H(\omega, \theta_F)\left(\tilde{\mathbf{\Phi}}_{\mathbf{U}}^{(N)} + \mu \mathbf{I}_N\right)^{-1} \mathbf{G}(\omega, \theta_F)}$$

where $G(\omega, \theta_F)$ denotes the front sound source steering vector

$$\mathbf{G}(\omega, \theta) = \begin{bmatrix} G_0(\omega, \theta) & \dots & G_{N-1}(\omega, \theta) \end{bmatrix}^T$$

[0092] I_N represents the N×N identity matrix, μ is a regularization parameter, and $\tilde{\mathbf{\Phi}}_{\mathbf{U}}^{(N)}$ denotes the normalized diffuse noise correlation matrix, which can be calculated from the joint acoustic and microphone responses as follows,

$$\tilde{\mathbf{\Phi}}_{\mathbf{U}}^{(N)} = \begin{bmatrix} \tilde{\Phi}_{U_0,U_0}^{(N)} & \dots & \tilde{\Phi}_{U_0,U_{N-1}}^{(N)} \\ \vdots & \ddots & \vdots \\ \tilde{\Phi}_{U_{N-1},U_0}^{(N)} & \dots & \tilde{\Phi}_{U_{N-1},U_{N-1}}^{(N)} \end{bmatrix}$$

with

$$\tilde{\Phi}_{U_m,U_n}^{(N)} = \frac{1}{2\pi}\int_0^{2\pi} G_m(\omega, \theta) G_n^*(\omega, \theta) d\theta$$

[0093] The directivity factor (DF) and the ratio of the front and back response (FBRR) of the SD beamformer are defined as follows:

$$\mathrm{DF}[dB] = 10\log_{10}\left(\frac{1}{2\pi}\int_0^{2\pi} \frac{\left|\mathbf{W}_{1,(:,1)}^H(\omega)\mathbf{G}(\omega, \theta_F)\right|^2}{\mathbf{W}_{1,(:,1)}^H(\omega)\tilde{\mathbf{\Phi}}_{\mathbf{U}}^{(N)}\mathbf{W}_{1,(:,1)}(\omega)} d\omega\right)$$

$$\mathrm{FBRR}[dB] = 10\log_{10}\left(\frac{\int_0^{2\pi}\left|\mathbf{W}_{1,(:,1)}^H(\omega)\mathbf{G}(\omega, \theta_F)\right|^2 d\omega}{\int_0^{2\pi}\left|\mathbf{W}_{1,(:,1)}^H(\omega)\mathbf{G}(\omega, \theta_B)\right|^2 d\omega}\right).$$

[0094] Whereas the DF is nearly constant with FSB filter length, the FBRR increases for higher filter lengths and approximately saturates for a length greater than or equal to 128. Note that the frequency-domain SD design is executed at $L_{FSB}/2$ frequencies that are uniformly distributed in the Nyquist interval, after which the frequency-domain FSB coefficients are transformed to length-L_{FSB} time-domain filters. Experiments have also shown a significant performance gap between the 2-mic configuration and other configurations, with greater than 2 microphones, both in terms of directivity and FBRR.

[0095] The BM in the fixed beamformer consists of a number of filter-and-sum beamformers that each operate on one particular subset of microphone signals. In this way, a number of noise reference signals is created, in which the power of the desired signal components is maximally reduced relative to the power of these components in the microphone signals. Typically, in an N-microphone configuration, N-1 noise references are created by designing N-1 different filter-

and-sum beamformers. However, in some cases it might be preferable to create fewer than N-1 noise references, which then leads to a reduction of the number of input signals $x_n(t)$ for the adaptive beamformer. In fact, in this embodiment we employ a BM consisting of only one filter-and-sum beamformer designed using the complete set of available microphone signals. In this way, the number of adaptive filters and hence the computational complexity of the adaptive beamformer can be considerably reduced.

[0096] In the context of the BM design, we consider the back sound source (if any) to be an undesired signal (which should be cancelled by the adaptive beamformer); hence the BM design reduces to a front-cancelling beamformer (FCB) design. Again, one of several different fixed beamformer design methods can be employed. In this embodiment, we use an array pattern synthesis method, different from existing methods.

[0097] In general, we can specify the frequency-domain FCB design at a set of angles $\{\theta_0,....,\theta_{M-1}\}$ by the following linear system of equations:

$$\underbrace{\begin{bmatrix} G_0^*(\omega,\theta_0) & \dots & G_{N-1}^*(\omega,\theta_0) \\ \vdots & \ddots & \vdots \\ G_0^*(\omega,\theta_{M-1}) & \dots & G_{N-1}^*(\omega,\theta_{M-1}) \end{bmatrix}}_{\bar{\mathbf{G}}^H(\omega)} \mathbf{W}_{1,(:,2)}(\omega) = \underbrace{\begin{bmatrix} P_0^*(\omega) \\ \vdots \\ P_{M-1}^*(\omega) \end{bmatrix}}_{\mathbf{P}^*(\omega)}$$

where $P_m(\omega)$, $m = 0,...,M-1$ denotes the desired response at frequency ω and angle θ_m. The least-squares (LS) optimal solution is then given by

$$\mathbf{W}_{1,(:,2)}(\omega) = \left[\bar{\mathbf{G}}(\omega)\bar{\mathbf{G}}^H(\omega)\right]^{-1}\bar{\mathbf{G}}(\omega)\mathbf{P}^*(\omega)$$

[0098] More specifically, to obtain an FCB design we should specify a zero response in the front direction and a non-zero response in any other direction. Preferably the latter direction should the back direction to avoid that the design would actually correspond to a front-back-cancelling beamformer design. As a consequence, the number of equations in the linear system of equations above is M=2, and the specification angles correspond to $\theta_0 = \theta_F$ and $\theta_1 = \theta_B$. Finally, the desired response vector is equal to $P^*(\omega) = [0,1]^H$.

[0099] With this design, the back response is indeed close to a unity response for most microphone configurations and filter length values. However, the front source response varies heavily according to the microphone configuration and filter length used. An important observation is that at least one microphone pair in an endfire configuration should preferably be included in the array to obtain a satisfactory power reduction of the front sound source component. Concerning the choice of the BM filter length, experiments show that there is no clear threshold effect - that is, the response in the front direction decreases with a nearly constant slope (provided an endfire microphone pair is included). As a consequence, the BM filter length should preferably be chosen according to the desired front sound source power reduction.

[0100] The design of the adaptive beamformer in an exemplary embodiment will now be described.

[0101] The adaptive beamformer in the block scheme may be implemented using a generalized sidelobe canceller (GSC) algorithm; a multi-channel Wiener filtering (MWF) algorithm; or any other adaptive algorithm. In this embodiment, we employ the speech-distortion-weighted multi-channel Wiener filtering (SDW-MWF) which includes the GSC and MWF as special cases. Details of this method can be found in S. Doclo, A. Spriet, J. Wouters, and M. Moonen ("Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction," Speech Commun., vol. 49, no. 7-8, pp. 636-656, Jul.-Aug. 2007, special Issue on Speech Enhancement).

[0102] The objective of the SDW-MWF is to jointly minimize the energy of the undesired components (B, I, N) and the distortion of the desired component (F) in the enhanced signal $Y_0(\omega)$. That is,

$$\min_{\mathbf{W}_2(\omega)} E\left\{\left|\mathbf{W}_2^H(\omega)\left[\mathbf{X}^{(B)}(\omega)+\mathbf{X}^{(I)}(\omega)+\mathbf{X}^{(N)}(\omega)\right]\right|^2\right\} + \frac{1}{\mu}E\left\{\left|X_0^{(F)}(\omega)-\mathbf{W}_2^H(\omega)\mathbf{X}^{(F)}(\omega)\right|^2\right\}$$

resulting in the adaptive beamformer estimate

$$\mathbf{W}_2(\omega) = \left[\mathbf{\Phi}_{\mathbf{x}}{}^{(F)}(\omega) + \mu\mathbf{\Phi}_{\mathbf{x}}{}^{(B,I,N)}(\omega)\right]^{-1}\mathbf{\Phi}_{\mathbf{x}}{}^{(F)}(\omega)\mathbf{e}_0$$

where $\mathbf{e}_0 \triangleq [1, 0, \ldots, 0]^T$ and the correlation matrices of the desired and undesired components in the adaptive beamformer input signal are defined as

$$\mathbf{\Phi}_{\mathbf{x}}{}^{(F)}(\omega) = E\left\{\left[\mathbf{X}^{(F)}(\omega)\right]\left[\mathbf{X}^{(F)}(\omega)\right]^H\right\}$$

$$\mathbf{\Phi}_{\mathbf{x}}{}^{(B,I,N)}(\omega) = E\left\{\left[\mathbf{X}^{(B)}(\omega) + \mathbf{X}^{(I)}(\omega) + \mathbf{X}^{(N)}(\omega)\right]\left[\mathbf{X}^{(B)}(\omega) + \mathbf{X}^{(I)}(\omega) + \mathbf{X}^{(N)}(\omega)\right]^H\right\}$$

[0103] The parameter μ can be tuned to trade off energy reduction of the undesired components versus distortion of the desired component. Several recursive implementations of the SDW-MWF filter estimate have been proposed, in which the adaptive SDW-MWF filter update is based on a generalized singular value decomposition (GSVD), a QR decomposition (QRD), a time-domain stochastic gradient method, or a frequency-domain stochastic gradient method. A common feature of these implementations is that the correlation matrices $\Phi_x{}^{(F)}(\omega)$ and $\Phi_x{}^{(B,I,N)}(\omega)$ are explicitly estimated before the SDW-MWF filter estimate is computed.

[0104] The signal-to-noise ratio (SNR) improvement provided by the SDW-MWF adaptive beamformer has been evaluated in a scenario with two localized sound sources: a front sound source consisting of a male speech signal (θ_F=0) and a localized interfering sound source consisting of a music signal (θ_I = 90 degrees).

[0105] The mean SNR at the microphones is equal to 10 dB. The fixed beamformer is implemented using a SD design for the FSB and a front-cancelling design for the BM, and an evaluation is done both for L_{FSB}=L_{BM}=64 and for L_{FSB}=L_{BM}=128. The adaptation of the SDW-MWF algorithm is based on a stochastic gradient frequency-domain implementation, and is controlled by a perfect (manual) voice activity detection (VAD). Two features of the SDW-MWF have been evaluated, namely:

1) the use of a feedforward filter $^{W}2,(1)^{(\omega)}$to include the fixed beamformer primary output signal $X_0(\omega)$ as an additional noise reference in the adaptive beamformer; and
2) the value of the SDW-MWF trade-off parameter $1/\mu$ (where $1/\mu = 0$ means no penalization of the desired component distortion).

[0106] Note that in case the desired component distortion is not penalized ($1/\mu$=0), the algorithm without a feedforward filter corresponds to the GSC algorithm, while the algorithm with a feedforward filter is not relevant due to an intolerable speech distortion. The evaluation has shown that the GSC algorithm as well as the SDW-MWF algorithm with a small trade-off parameter ($1/\mu = 0.01$) are well suited for the reduction of the localized interfering sound source power. Moreover, there appears to be no significant influence of the number of microphones and the FSB and BM filter lengths on the adaptive beamformer performance.

[0107] The design of the Adaptive Spectral Attenuation process in an exemplary embodiment will now be described.

[0108] The adaptive spectral attenuation block is included in the structure with the aim of reducing the diffuse noise energy in the primary adaptive beamformer output signal. To this end, the short-term magnitude spectra of the reference microphone signal, $|U_0(\omega_k,l)|$, and the primary and secondary adaptive beamformer output signals, $|Y_0(\omega_k,l)|$ and $|Y_1(\omega_k,l)|$, are estimated by means of a Discrete Fourier transform (DFT), with k and l denoting the DFT frequency bin and time frame indices. An instantaneous spectral gain function is then calculated as follows,

$$G_{\text{inst}}(\omega_k, l) = \frac{|U_0(\omega_k, l)| - \beta_n\hat{C}(\omega_k, l)|\hat{Y}_1(\omega_k, l)|}{|\hat{Y}_0(\omega_k, l)| + \varepsilon}$$

where the subtraction factor $\beta_n \in [0,1]$ determines the amount of spectral attenuation and the regularization factor ε is a small constant which prevents division by zero. Since the secondary adaptive beamformer output signal $Y_1(\omega)$ is equal

to the noise reference $X_1(\omega)$ at the output of the fixed beamformer, a spectral coherence function $C(\omega_k,l)$ that relates the magnitude spectra of the diffuse noise components in the primary and secondary fixed beamformer output signals needs to be estimated and taken into account in the equation. The instantaneous gain function of the equation is then lowpass filtered and clipped, before being applied to the speech estimate, that is,

$$G_{\mathrm{lp}}(\omega_k, l) = (1-\alpha)G_{\mathrm{lp}}(\omega_k, l-1) + \alpha G_{\mathrm{inst}}(\omega_k, l)$$

$$G(\omega_k, l) = \max\{G_{\mathrm{lp}}(\omega_k, l), \xi_n\}$$

$$|Z(\omega_k, l)| = G(\omega_k, l)|Y_0(\omega_k, l)|$$

where α denotes the lowpass filter pole and ξ_n =1 - β_n is the clipping level. The enhanced signal magnitude spectrum $|Z(\omega_k,l)|$ is subsequently transformed back to the time domain by applying an inverse DFT (IDFT), and by using the phase spectrum of the primary adaptive beamformer output signal $Y_0(\omega_k,l)$.

[0109] An exemplary use of the embodiment in an Acoustic Zoom (AZ) application will now be described.

1) Specification of the time-varying gain factors: In the AZ application, the aim is to keep the level of the undesired sound sources constant, while the level of the desired sound sources should adapt to the camera zoom state. As a consequence, we should set the gain factors for the localized interfering sound source and the diffuse noise as follows,

$$g_I(t) \equiv 1$$

$$g_N(t) \equiv 1$$

[0110] From preliminary results with the above zoom-in trajectory for the front sound source level, it was noted that a perceptually better trajectory could be designed. More particularly, a faster level increase at the start of the zoom-in operation would be desired, eventually converging to the same final level at close-up. A perceptually more attractive level trajectory was found to be

$$g_F(t) = 1 + \frac{2^{d_{\mathrm{zoom}}} - 1}{\sqrt{1.2 d_{\mathrm{zoom}}}} \sqrt{1.2 v_{\mathrm{zoom}} t}, \quad 0 \le t \le \frac{d_{\mathrm{zoom}}}{v_{\mathrm{zoom}}}$$

[0111] Concerning the specification of the back sound source gain factor, several possibilities exist. A first possibility is to regard the back sound source as an undesired sound source, in which case its level should remain constant. However, since the back sound source is typically very close to the camera, its level should often be reduced to obtain an acceptable balance between the back sound source and the other sound sources. A second possibility is to have the back sound source gain factor follow the inverse trajectory of the front sound source gain factor, possibly combined with a fixed back sound source level reduction. While such an inverse level trajectory would obviously make sense from a physical point of view, it may be perceived somewhat too artificial, since the front sound source level change is then supported by visual cues, while the back sound source level change is not.

[0112] Experiments have been performed to demonstrate the performance of the AZ algorithm. In both experiments, the front sound source is a male speech signal corresponding to a camera recording that consists of a far shot phase (5 s), a zoom-in phase (10 s), and a close-up phase (11 s). In addition, the sound field consists of diffuse babble noise and a localized interfering music source at θ_I = 90 deg. In the first simulation, no back sound source is present, while in the second simulation, a female speech signal is present in the back direction (θ_B = 180 deg).

[0113] A 3-microphone array was used, employing microphones 1,3, and 4 as indicated in Fig. 1. The fixed beamformer consists of a superdirective FSB and a single-noise-reference front-cancelling BM, both a with filter length of 64. The adaptive beamformer is calculated using a GSC algorithm and has a filter length of 128. The desired AZ effect consists

in keeping the level of the undesired sound sources (including the back sound source in the second simulation) unaltered, while increasing the level of the front sound source during the zoom-in phase, according to the perceptually optimal trajectory defined above.

[0114] In these embodiments the values of the re-weighting factors were determined empirically in advance, rather than at run-time (as described previously above).

[0115] As will be apparent to those skilled in the art, the performance of the method depends in part upon the accuracy to which the reweighting factors can be estimated. The greater the accuracy, the better the performance of the manipulation will be.

[0116] Fig. 4 is a flowchart summarising a method according to an embodiment. In step 410, audio signals 6 are received from the microphones 4. In step 420, the desired gain factors 8 are input. In step 430, the auxiliary signal generator generates the auxiliary signals. In step 440, the scaling coefficient calculator 30 calculates the scaling coefficients, a(t). Finally, in step 450, the audio synthesis unit 20 applies the scaling coefficients to the generated auxiliary signals and reference audio signals, to synthesise output audio signals 40.

[0117] While the invention has been illustrated and described in detail in the drawings and foregoing description, such illustration and description are to be considered illustrative or exemplary and not restrictive; the invention is not limited to the disclosed embodiments.

[0118] For example, it is possible to operate the invention in an embodiment wherein different blocks are used to generate the auxiliary signals. The exemplary blocks described above (fixed or adaptive beamforming, or adaptive spectral modification) can be replaced or supplemented by other methods. Essentially, the auxiliary signal calculation should be such that it exploits the diversity of the individual sound sources in the sound scene. When multiple microphones are used, then exploiting spatial diversity is often the most straightforward option - and this is exploited by the beamformers in the embodiments described above. However, different kinds of diversity could equally be exploited, for example: diversity in the time domain (if not all of the sound sources are concurrently active); diversity in statistics (which could lead to the use of Wiener filtering, independent component analysis, and so on); or diversity in the degree of (non-) stationarity. The optimal choice of auxiliary signal generator will vary according to the application and the characteristics of the audio environment.

[0119] The ordering of the blocks described in embodiments herein and shown in the drawings is also not limiting on the scope of the invention. Blocks may be eliminated, re-ordered or duplicated.

[0120] Likewise, although the embodiments described herein have concentrated on monaural or stereo implementation, the invention can of course be implemented with a greater number of audio output signals than just one or two. Those skilled in the art will be readily able to generalise from the description above, to provide an arbitrary number of desired outputs. This may be useful, for example, for multi-channel or surround-sound audio applications.

[0121] Other variations to the disclosed embodiments can be understood and effected by those skilled in the art in practicing the claimed invention, from a study of the drawings, the disclosure, and the appended claims. In the claims, the word "comprising" does not exclude other elements or steps, and the indefinite article "a" or "an" does not exclude a plurality. A single processor or other unit may fulfil the functions of several items recited in the claims. The mere fact that certain measures are recited in mutually different dependent claims does not indicate that a combination of these measured cannot be used to advantage. A computer program may be stored/distributed on a suitable medium, such as an optical storage medium or a solid-state medium supplied together with or as part of other hardware, but may also be distributed in other forms, such as via the Internet or other wired or wireless telecommunication systems. Any reference signs in the claims should not be construed as limiting the scope.

Claims

1. An audio-processing device comprising:

an audio input, for receiving one or more audio signals detected at respective microphones, each of the audio signals comprising a mixture of a plurality of components, each component corresponding to a sound source;
a control input, for receiving, for each sound source, a desired gain factor associated with the source, by which it is desired to amplify the corresponding component;
an auxiliary signal generator, adapted to generate at least one auxiliary signal from the one or more audio signals, the at least one auxiliary signal comprising a different mixture of the components as compared with a reference one of the one or more audio signals;
a scaling coefficient calculator, adapted to calculate a set of scaling coefficients in dependence upon the desired gain factors and upon parameters of the different mixture, each scaling coefficient associated with one of the at least one auxiliary signal and optionally the reference audio signal; and
an audio synthesis unit, adapted to synthesize an output audio signal by applying the scaling coefficients to the

at least one auxiliary signal and optionally the reference audio signal and to combine the results,

wherein the scaling coefficients are calculated from the desired gain factors and the parameters of the different mixture such that the synthesized output signal provides the desired gain factor for each component.

2. A handheld personal electronic device comprising a plurality of microphones; and
the audio processing device of claim 1.

3. The mobile or handheld device of claim 2, wherein the microphones are omni-directional microphones.

4. A method of processing audio signals comprising:

receiving one or more audio signals detected at respective microphones, each of the audio signals comprising a mixture of a plurality of components, each component corresponding to a sound source;
receiving, for each sound source, a desired gain factor associated with the source, by which it is desired to amplify the corresponding component;
generating at least one auxiliary signal from the one or more audio signals, the at least one auxiliary signal comprising a different mixture of the components as compared with a reference one of the one or more audio signals;
calculating a set of scaling coefficients in dependence upon the desired gain factors and upon parameters of the different mixture, each scaling coefficient associated with one of the at least one auxiliary signal and optionally the reference audio signal; and
synthesizing an output audio signal by applying the scaling coefficients to the at least one auxiliary signal and optionally the reference audio signal and combining the results,

wherein the scaling coefficients are calculated from the desired gain factors and the parameters of the different mixture such that the synthesized output signal provides the desired gain factor for each component.

5. The method of claim 4, wherein the parameters of the different mixture are reweighting factors, which relate the levels of the components in the at least one auxiliary signal to their respective levels in the reference audio signal.

6. The method of claim 5, wherein:

the desired gain factors; the reweighting factors and the scaling coefficients are related by a linear system equations; and
the step of calculating the set of scaling coefficients comprises solving the system of equations.

7. The method of any of claims 4 to 6, wherein the at least one auxiliary signal is a linear combination of any of:

one or more of the audio signals;
one or more temporally shifted versions of the audio signals; and
one or more filtered versions of the audio signals.

8. The method of any of claims 4 to 7, wherein the at least one auxiliary signal is generated by at least one of:

fixed beamforming;
adaptive beamforming; and
adaptive spectral modification.

9. The method of any of claims 4 to 8, wherein:

a first auxiliary signal is generated by a first method;
a second auxiliary signal is generated by a second, different method;
and
the second auxiliary signal is generated based on an output of the first method.

10. The method of any of claims 4 to 9, comprising:

synthesizing a first output audio signal by applying scaling coefficients to a first reference audio signal and at least one first auxiliary signal and combining the results; and
synthesizing a second output audio signal by applying scaling coefficients to a second, different reference audio signal and at least one second auxiliary signal and combining the results.

11. The method of any of claims 4 to 10, wherein the sound sources comprise one or more localised sound sources and a diffuse noise field.

12. The method of any of claims 4 to 11, wherein the desired gain factors are time-varying.

13. The method of any of claims 4 to 12,
wherein the desired gain factors are chosen in dependence upon the state of a visual zoom function.

14. A computer program comprising computer program code means adapted to perform all the steps of any of claims 4 to 13 when said program is run on a computer.

15. A computer program as claimed in claim 14 embodied on a computer readable medium.

Amended claims in accordance with Rule 137(2) EPC.

1. An audio-processing device comprising:

an audio input, for receiving one or more audio signals (6) detected at respective microphones (4), each of the audio signals comprising a mixture of a plurality of components, each component corresponding to a sound source;
a control input, for receiving, for each sound source, a desired gain factor (8) associated with the source, by which it is desired to amplify the corresponding component;
an auxiliary signal generator (10, 210, 212, 214), adapted to generate at least one auxiliary signal from the one or more audio signals, the at least one auxiliary signal comprising a different mixture of the components as compared with a reference one of the one or more audio signals, wherein the levels of the components in the at least one auxiliary signal are related to their respective levels in the reference audio signal by known reweighting factors;
a scaling coefficient calculator (30), adapted to calculate a set of scaling coefficients in dependence upon the desired gain factors and upon the reweighting factors, each scaling coefficient associated with one of the at least one auxiliary signal and optionally the reference audio signal; and
an audio synthesis unit (20, 220, 320), adapted to synthesize an output audio signal (40) by applying the scaling coefficients to the at least one auxiliary signal and optionally the reference audio signal and to combine the results,
wherein the scaling coefficients are calculated from the desired gain factors and the reweighting factorssuch that the synthesized output signal (40) provides the desired gain factor for each component.

2. A handheld personal electronic device comprising
a plurality of microphones (4); and
the audio processing device of claim 1.

3. The mobile or handheld device of claim 2, wherein the microphones (4) are omni-directional microphones.

4. A method of processing audio signals comprising:

receiving (410) one or more audio signals detected at respective microphones, each of the audio signals comprising a mixture of a plurality of components, each component corresponding to a sound source;
receiving (420), for each sound source, a desired gain factor associated with the source, by which it is desired to amplify the corresponding component;
generating (430) at least one auxiliary signal from the one or more audio signals, the at least one auxiliary signal comprising a different mixture of the components as compared with a reference one of the one or more audio signals, wherein the levels of the components in the at least one auxiliary signal are related to their respective levels in the reference audio signal by known reweighting factors;
calculating (440) a set of scaling coefficients in dependence upon the desired gain factors and upon the re-

weighting factors, each scaling coefficient associated with one of the at least one auxiliary signal and optionally the reference audio signal; and
synthesizing (450) an output audio signal by applying the scaling coefficients to the at least one auxiliary signal and optionally the reference audio signal and combining the results,
wherein the scaling coefficients are calculated from the desired gain factors and the reweighting factors such that the synthesized output signal provides the desired gain factor for each component.

5. The method of claim 4, wherein:

the desired gain factors; the reweighting factors and the scaling coefficients are related by a linear system equations; and
the step of calculating the set of scaling coefficients comprises solving the system of equations.

6. The method of claim 4 or 5, wherein the at least one auxiliary signal is a linear combination of any of:

one or more of the audio signals;
one or more temporally shifted versions of the audio signals; and
one or more filtered versions of the audio signals.

7. The method of any of claims 4 to 6, wherein the at least one auxiliary signal is generated by at least one of:

fixed beamforming;
adaptive beamforming; and
adaptive spectral modification.

8. The method of any of claims 4 to 7, wherein:

a first auxiliary signal is generated by a first method;
a second auxiliary signal is generated by a second, different method; and
the second auxiliary signal is generated based on an output of the first method.

9. The method of any of claims 4 to 8, comprising:

synthesizing a first output audio signal by applying scaling coefficients to a first reference audio signal and at least one first auxiliary signal and combining the results; and
synthesizing a second output audio signal by applying scaling coefficients to a second, different reference audio signal and at least one second auxiliary signal and combining the results.

10. The method of any of claims 4 to 9, wherein the sound sources comprise one or more localised sound sources and a diffuse noise field.

11. The method of any of claims 4 to 10, wherein the desired gain factors are time-varying.

12. The method of any of claims 4 to 11,
wherein the desired gain factors are chosen in dependence upon the state of a visual zoom function.

13. A computer program comprising computer program code means adapted to perform all the steps of any of claims 4 to 12 when said program is run on a computer.

14. A computer program as claimed in claim 13 embodied on a computer readable medium.

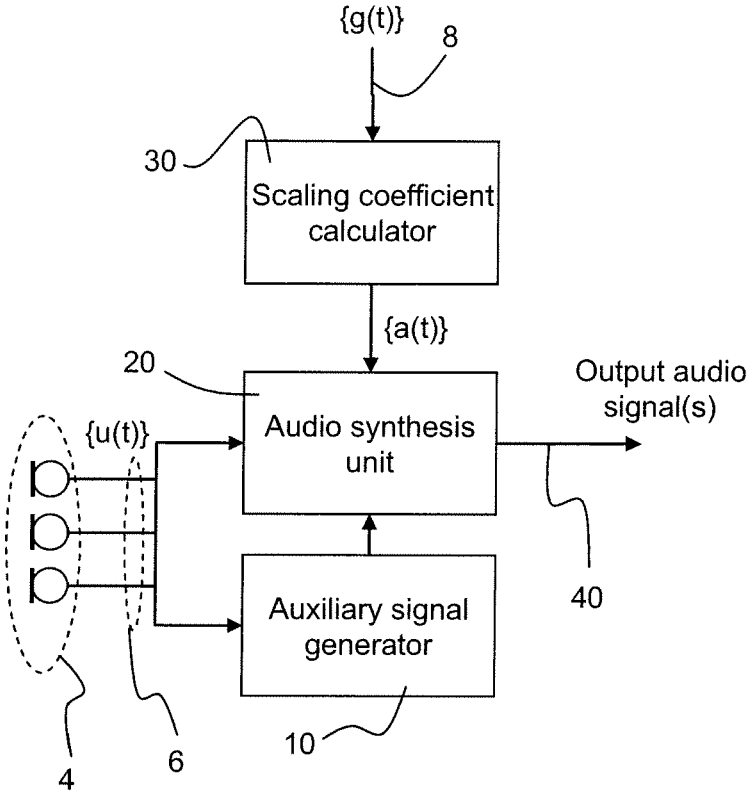
{g(t)}
8
30
Scaling coefficient calculator
{a(t)}
20
Output audio signal(s)
{u(t)}
Audio synthesis unit
Auxiliary signal generator
40
4
6
10

FIG 1

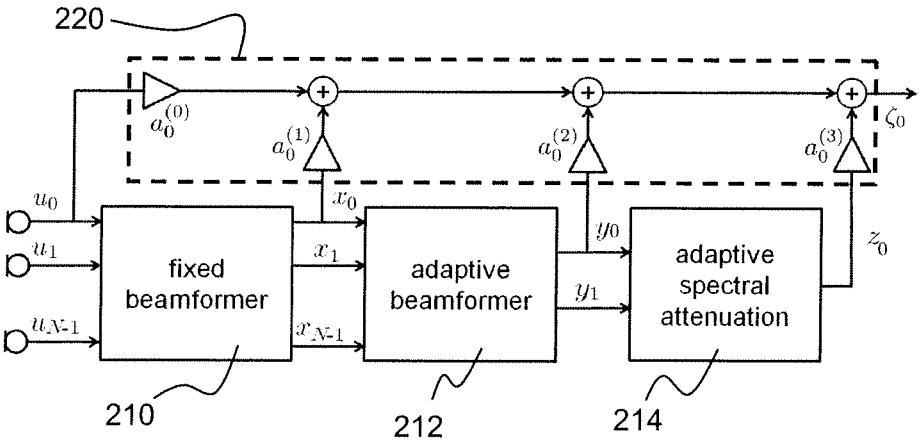
220
$a_0^{(0)}$
$a_0^{(1)}$
$a_0^{(2)}$
$a_0^{(3)}$
ζ_0
u_0
u_1
u_{N-1}
x_0
x_1
x_{N-1}
fixed beamformer
adaptive beamformer
y_0
y_1
adaptive spectral attenuation
z_0
210
212
214

FIG 2

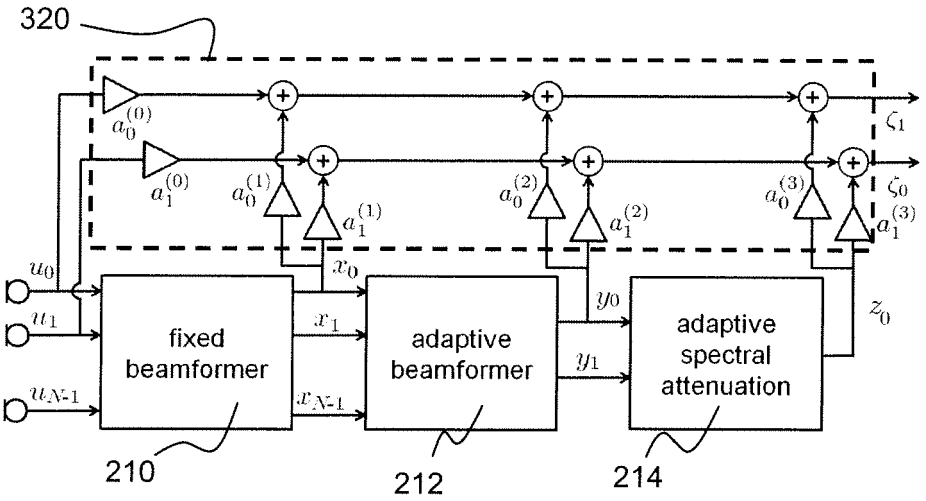
320
$a_0^{(0)}$
$a_1^{(0)}$
$a_0^{(1)}$
$a_1^{(1)}$
$a_0^{(2)}$
$a_1^{(2)}$
$a_0^{(3)}$
$a_1^{(3)}$
ζ_1
ζ_0
u_0
u_1
u_{N-1}
x_0
x_1
x_{N-1}
y_0
y_1
z_0
fixed beamformer
adaptive beamformer
adaptive spectral attenuation
210
212
214

FIG 3

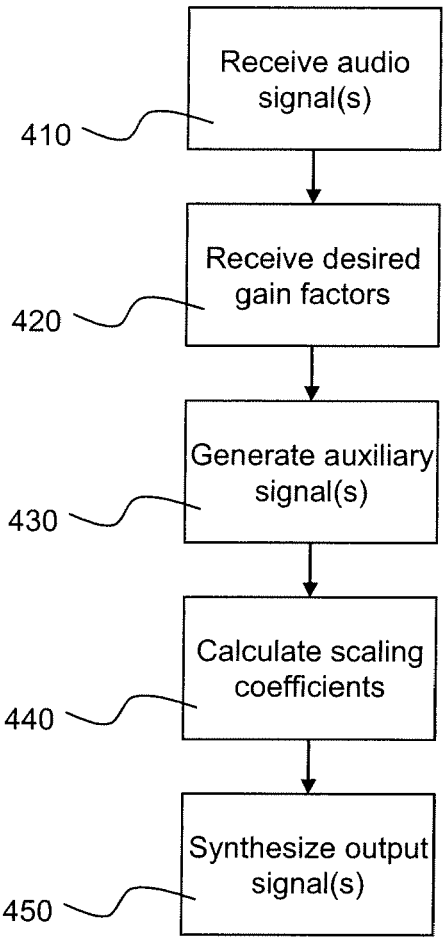
Receive audio signal(s)
410
Receive desired gain factors
420
Generate auxiliary signal(s)
430
Calculate scaling coefficients
440
Synthesize output signal(s)
450

FIG 4

Europäisches Patentamt
European Patent Office
Office européen des brevets

EUROPEAN SEARCH REPORT

Application Number
EP 10 27 5102

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (IPC)
A	JANG INSEON ET AL: "An Object-based 3D Audio Broadcasting System for Interactive Services", AES CONVENTION 118; MAY 2005, AES, 60 EAST 42ND STREET, ROOM 2520 NEW YORK 10165-2520, USA, 1 May 2005 (2005-05-01), XP040507192, * page 4 - page 7 *	1-15	INV. H04R3/00 G10L21/02
A	GB 2 353 193 A (YAMAHA CORP [JP]) 14 February 2001 (2001-02-14) * abstract; figures 1-7 *	1-15	
A	EP 2 131 610 A1 (STARKEY LAB INC [US]) 9 December 2009 (2009-12-09) * abstract; figure 8 *	1-15	

TECHNICAL FIELDS SEARCHED (IPC)
H04R
G10L
H04S

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
Munich	16 February 2011	Righetti, Marco

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)

ANNEX TO THE EUROPEAN SEARCH REPORT ON EUROPEAN PATENT APPLICATION NO. EP 10 27 5102

This annex lists the patent family members relating to the patent documents cited in the above-mentioned European search report. The members are as contained in the European Patent Office EDP file on
The European Patent Office is in no way liable for these particulars which are merely given for the purpose of information.

16-02-2011

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
GB 2353193 A	14-02-2001	JP 2001069597 A	16-03-2001
		US 7162045 B1	09-01-2007
EP 2131610 A1	09-12-2009	AT 478525 T	15-09-2010
		DK 2131610 T3	27-09-2010
		US 2009296944 A1	03-12-2009

EPO FORM P0459

For more details about this annex : see Official Journal of the European Patent Office, No. 12/82

REFERENCES CITED IN THE DESCRIPTION

This list of references cited by the applicant is for the reader's convenience only. It does not form part of the European patent document. Even though great care has been taken in compiling the references, errors or omissions cannot be excluded and the EPO disclaims all liability in this regard.

Non-patent literature cited in the description

- **S. DOCLO ; M. MOONEN.** Superdirective beamforming robust against microphone mismatch. *IEEE Trans. Audio Speech Lang. Process.,* February 2007, vol. 15 (2), 617-631 **[0091]**
- **S. DOCLO ; A. SPRIET ; J. WOUTERS ; M. MOONEN.** Frequency-domain criterion for the speech distortion weighted multichannel wiener filter for robust noise reduction. *Speech Commun.,* July 2007, vol. 49, 636-656 **[0101]**